经验分布的一致收敛及其在 Non Free Lunch 定理的极限情况中的应用

王俊彦
浙江经济职业技术学院

日期：January 8, 2024

摘　要

No Free Lunch(NFL) 定理是统计学习理论的一个重要结果 (Wolpert, 1992, 1996, 2002)，依据贝叶斯建模可以推得损失/效用函数的期望与预测函数的假设空间的选取有关，若认为真实的预测函数空间是不可知的，则任意选择的假设函数空间都不一定得到最优的损失函数的期望。

本文对 NFL 定理的极限情况进行分析，利用分布的一致收敛性，即 Glivenko-Cantelli 定理 (VI, 1933; Cantelli, 1933; Dvoretzky et al., 1956; 韦来生, 2008; 茆诗松 等, 2006) 的一种局部形式得到——在一定情况下的确定性与非确定性预测问题中，当样本量趋于无穷大损失/效用函数的期望与假设函数空间的具体选择无关。此项工作的一个副产物是利用本文得出的分布的一致收敛性的局部形式可以推得分布的总变差（total variation）一致收敛性。此前该性质一般是认为不存在的 (Devroye et al., 1990)。

1　经验分布一致收敛的三种形式

Glivenko-Cantelli 定理是数理统计理论的基本定理之一，可以描述为：

定理 1.1. 定义在概率空间 $(X, \mathcal{A}, \mathcal{P})$ 上的累积分布函数 $F(x)$，及其经验累积分布 $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \le x}$:

$$\lim_{n\to\infty} \mathcal{P}(\sup_x |F_n(x) - F(x)| > 0) = 0 \tag{1}$$

为了方便后续 No Free Lunch 定理中的讨论，这里还给出 Glivenko-Cantelli 定理的一个推论：

推论 1.2. 定义在概率空间 $(X, \mathcal{A}, \mathcal{P}), X \subset \mathbb{R}$ 上的累积分布函数 $F(x)$，及其经验累积分布 $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{x_i \le x}$，满足 $\forall \Delta > 0, \varepsilon > 0$:

$$\lim_{n\to\infty} \mathcal{P}\left(\sup_x \left|D_\Delta^+ F_n(x) - D_\Delta^+ F(x)\right| > \varepsilon\right) = 0 \tag{2}$$

其中，$D_\Delta^+ f(x) = \frac{f(x+\Delta)-f(x)}{\Delta}$.

证明.

$$\begin{aligned}&\sup_x \left|D^+_\Delta F_n(x) - D^+_\Delta F(x)\right| \\ &= \sup_x \left|\frac{F_n(x+\Delta)-F_n(x)}{\Delta} - \frac{F(x+\Delta)-F(x)}{\Delta}\right| \\ &\le \frac{1}{|\Delta|}\sup_x |F_n(x+\Delta)-F(x+\Delta)| + \sup_x |F_n(x)-F(x)|\end{aligned} \tag{3}$$

令 $\epsilon = \frac{1}{2}|\Delta|\varepsilon$，且 $\sup_x |F_n(x+\Delta)-F(x+\Delta)| \le \epsilon, \sup_x |F_n(x)-F(x)| \le \epsilon$，则

$$\sup_x \left|D^+_\Delta F_n(x) - D^+_\Delta F(x)\right| \le \varepsilon \tag{4}$$

由 Glivenco-Cantelli 定理知对任意 $\epsilon > 0$ 有

$$\lim_{n\to\infty} \mathcal{P}(\sup_x |F_n(x)-F(x)| \le \epsilon) = 1 - \lim_{n\to\infty} \mathcal{P}(\sup_x |F_n(x)-F(x)| > \epsilon) = 1 \tag{5}$$

注意到：

$$\begin{aligned}&\lim_{n\to\infty} \mathcal{P}(\sup_x |F_n(x)-F(x)| > \epsilon) = 0 \\ &\lim_{n\to\infty} \sup_x |F_n(x+\Delta)-F(x+\Delta)| > \epsilon) = 0 \\ \Longrightarrow &\lim_{n\to\infty} \sup_x |F_n(x+\Delta)-F(x+\Delta)| > \epsilon \cup \sup_x |F_n(x)-F(x)| > \epsilon) = 0 \\ \Longrightarrow &\lim_{n\to\infty} \sup_x |F_n(x+\Delta)-F(x+\Delta)| \le \epsilon \cap \sup_x |F_n(x)-F(x)| \le \epsilon) = 1\end{aligned} \tag{6}$$

故 $\forall \varepsilon > 0$：

$$\begin{aligned}&\lim_{n\to\infty} \mathcal{P}(\sup_x \left|D^+_\Delta F_n(x) - D^+_\Delta F(x)\right| \le \varepsilon) \\ &\ge \lim_{n\to\infty} \mathcal{P}(\sup_x |F_n(x)-F(x)| \le \epsilon, \sup_x |F_n(x+\Delta)-F(x+\Delta)| \le \epsilon) \\ &= 1\end{aligned} \tag{7}$$

故得证。 □

类似的可以得到：

推论 1.3. 定义在概率空间 $(X, \mathcal{A}, \mathcal{P}), X \subset \mathbb{R}$ 上的累积分布函数 $F(x)$, 及其经验累积分布 $F_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i \le x}$，满足 $\forall \Delta > 0, \varepsilon > 0$:

$$\lim_{n\to\infty} \mathcal{P}\left(\sup_x |D_\Delta F_n(x) - D_\Delta F(x)| > \varepsilon\right) = 0 \tag{8}$$

其中，$D_\Delta f(x) = \frac{f(x+\Delta)-f(x-\Delta)}{2\Delta}$.

由该定理可以推得一个关于总变差 (total variation) 距离的结论：

命题 1.4. 定义在概率空间 $(X, \mathcal{A}, \mathcal{P}), X \subset \mathbb{R}$，且 X 有界的概率测度函数 $P : \mathcal{A} \mapsto [0,1]$, 及其经验概率测度 $P_n(x) = \frac{1}{n}\sum_{i=1}^n \mathbf{1}_{x_i \in A}$，满足 $\forall \upsilon > 0$ 有：

$$\lim_{n\to\infty} \mathcal{P}(\sup_A |P^n(A) - P(A)| > \upsilon) = 0 \tag{9}$$

证明. 首先以半径为 ϵ 的小球覆盖 $A \subset A_\epsilon = \bigcup_{i=1}^N B^\epsilon(x_i)$。这是由于 A 完全有界，所以 $A \subset \bar{A}$

总可以由 $N<\infty$ 个小球覆盖，故

$$
\begin{aligned}
&|P^n(A)-P(A)-(P^n(A_\epsilon)-P(A_\epsilon))| \leq \int_{A_\epsilon} dx \leq N\text{vol}(B^\epsilon) \\
\Longrightarrow &|P^n(A)-P(A)| \leq |P^n(A_\epsilon)-P(A_\epsilon)| + N\text{vol}(B^\epsilon) \\
\Longrightarrow &|P^n(A)-P(A)| \leq \sum_{i=1}^{N} |P^n(B_\epsilon(x_i) - P(B_\epsilon(x_i)| + N\text{vol}(B^\epsilon) \\
\Longrightarrow &|P^n(A)-P(A)| \leq 2\epsilon \sum_{i=1}^{N} |D_\epsilon P^n(B_\epsilon(x_i) - D_\epsilon P(B_\epsilon(x_i)| + N\text{vol}(B^\epsilon) \\
\Longrightarrow &\sup_A |P^n(A)-P(A)| \leq \sup_A 2\epsilon \sum_{i=1}^{N} |D_\epsilon P^n(B_\epsilon(x_i) - D_\epsilon P(B_\epsilon(x_i)| + N\text{vol}(B^\epsilon) \\
\Longrightarrow &\sup_A |P^n(A)-P(A)| \leq 2\epsilon N_{max} \sup_x |D_\epsilon P^n(B_\epsilon(x_i) - D_\epsilon P(B_\epsilon(x_i)| + N_{max}\text{vol}(B^\epsilon)
\end{aligned}
\tag{10}
$$

其中 N_{max} 为以半径为 ϵ 小球覆盖任意 $A\in\mathcal{A}$ 的最大数。由于 $A\subset X$，X 存在有限覆盖，故任意 A 也存在有限覆盖。

由推论1.3，知 $\forall\epsilon>0,\varepsilon>0$ 有：

$$
\lim_{n\to\infty} \mathcal{P}(\sup_A |P^n(A)-P(A)| > N_{max}(2\epsilon\varepsilon + \text{vol}(B^\epsilon))) = 0 \tag{11}
$$

令 $\upsilon = N_{max}(2\epsilon\varepsilon+\text{vol}(B^\epsilon))$, 得证。 □

2 关于 No Free Lunch 定理的讨论

学习理论中的 No Free Lunch 定理的标准形式为：

定理 2.1. *设 C 为学习损失（或效用）；数据的集合为 $D_m=\{(x_1,y_1),(x_2,y_2),...(x_m,y_m)\}$，其中 x_i 为输入，y_i 为输出；真实函数空间的表示输入输出关系的函数为 f，其服从以 $P(f)$ 为概率的分布，解空间的表示输入输出关系的函数为 h，其服从以 $P(h)$ 为概率的分布，则：*

$$
E(C|D_m) = \sum_{f,h} \mathcal{E}(C|f,h,D_m)p(h|D_m)p(f|D_m) \tag{12}
$$

其中 $\mathcal{E}(C|f,h,D_m)$ 是关于 C 的条件期望。

可见，该定理是统计学习问题的贝叶斯描述。该定理常常可以悲观的理解为：由于 $p(f|D_m)$ 的分布未知，因此不能保证由学习算法所决定的 $p(h|D_m)$ 与 $p(f|D_m)$ 加权求和能得到最大的或最小的期望值。但是注意到这是在 m 有限的情形。如下将对 $m\to\infty$ 的情形进行分析。

确定性的输出 首先我们讨论一种常见的简单情形，即 $\exists f^*: X\mapsto y, y=f^*(x)$。对于学习过程我们假设：

1. $h\in\mathcal{H},\mathcal{H}$ 为假设空间，$f\in\mathcal{F},\mathcal{F}$ 为真实空间;
2. 且 $\mathcal{H}^{D_m}=\{h:h(x)=y,\forall(x,y)\in D_m, h\in\mathcal{H}\}$, $\mathcal{F}^{D_m}=\{f:f(x)=y,\forall(x,y)\in D_m\}$。

显然，

命题 2.2.

$$\mathcal{H} \supseteq \mathcal{H}^{D_1} \supseteq \mathcal{H}^{D_2} \supseteq ... \\ \mathcal{F} \supseteq \mathcal{F}^{D_1} \supseteq \mathcal{F}^{D_2} \supseteq ... \tag{13}$$

更进一步地，我们可以得到：

命题 2.3. 若设 $(X, \mathcal{A}, P)$ 是一个概率测度空间，且其累积概率分布 $F(x)$ 是 *Lipchitz* 的，即

$$\forall x, x' \in X, |F(x) - F(x')| \leq L|x - x'|, L < \infty \tag{14}$$

且 D_m 中元素各不相同，则如下等式以概率 *1* 成立：

$$\lim_{m\to\infty} \mathcal{P}(f(x) \neq f^*(x)|D_m) = 0, \lim_{m\to\infty} \mathcal{P}(h(x) \neq f^*(x)|D_m) = 0 \tag{15}$$

证明. 注意到 $\mathcal{P}(f(x) = f^*(x), x \in D_m|D_m) = \mathcal{P}(x \in D_m)$，则

$$\begin{aligned} &\mathcal{P}(f(x) = f^*(x)|D_m) \\ &= \mathcal{P}(f(x) = f^*(x), x \in D_m|D_m) + \mathcal{P}(f(x) = f^*(x), x \notin D_m|D_m) \\ &\geq \mathcal{P}(x \in D_m) \end{aligned} \tag{16}$$

有

$$\begin{aligned} &\mathcal{P}(f(x) \neq f^*(x)|D_m) \\ &= 1 - \mathcal{P}(f(x) = f^*(x)|D_m) \\ &\leq 1 - \mathcal{P}(x \in D_m) \\ &= 1 - \sum_{j=1}^{m} P(x = x_j), \text{注意到 } x_j \text{ 各不相同} \\ &= 1 - \sum_{j=1}^{m} \lim_{\Delta\to 0} \frac{F(x_j + \Delta) - F(x_j)}{\Delta}\Delta \end{aligned} \tag{17}$$

考虑如下构造，对 X 作间距为 Δ 的等分为 $x_1, x_2, ...$，则：

$$\lim_{\Delta\to 0} \sum_{i=1}^{|X|/\Delta} \lim_{\Delta\to 0} \frac{F(x_i + \Delta) - F(x_i)}{\Delta}\Delta = \int_X p(x)dx = P(x \in X) = 1 \tag{18}$$

注意到上述求和去掉了 $\{x_i\}$ 中最后一个点。

由推论1.2知，当 $m \to \infty$，对任意 $\forall \epsilon > 0, \forall [x, x+\epsilon) \subset X$, 若 $F(x+\epsilon) - F(x) > 0$ 则以概率 1 有 $F^m(x+\epsilon) - F^m(x) > 0$。因此 $\forall [x_i, x_i + \epsilon) \subset \mathcal{A}$，$\mathcal{P}(\exists x' \in D_m, 0 < x'_i - x_i < \epsilon) = 1$。对 $x_1, x_2, ...$ 取 $x'_1, x'_2, ...$, 使得 $0 < x'_i - x_i < \frac{\Delta}{2}$, 则由 Lipchitz 条件知，以概率 1 有：

$$\begin{aligned} &\lim_{m\to\infty} \left| \sum_{i=1}^{|X|/\Delta} \lim_{\Delta\to 0} \frac{F(x_i + \Delta) - F(x_i)}{\Delta}\Delta - \sum_{i=1}^{|X|/\Delta} \lim_{\Delta\to 0} \frac{F(x'_i + \Delta) - F(x'_i)}{\Delta}\Delta \right| \\ &\leq \lim_{\Delta\to 0} L\frac{|X|}{\Delta}\Delta^2 = 0 \end{aligned} \tag{19}$$

又 $\{x'_i\} \subset D_m$，故

$$\sum_{i=1}^{m} \lim_{\Delta\to 0} \frac{F(x_j + \Delta) - F(x_j)}{\Delta}\Delta \geq \sum_{i=1}^{|X|/\Delta} \lim_{\Delta\to 0} \frac{F(x'_i + \Delta) - F(x'_i)}{\Delta}\Delta \tag{20}$$

因此以概率 1 有：

$$
\begin{aligned}
&\lim_{m\to\infty}\mathcal{P}(f(x)\neq f^*(x)|D_m)\\
&\leq 1-\lim_{m\to\infty}\sum_{i=1}^{|X|/\Delta}\lim_{\Delta\to 0}\frac{F(x_i'+\Delta)-F(x_i')}{\Delta}\Delta\\
&=1-\lim_{\Delta\to 0}\sum_{i=1}^{|X|/\Delta}\lim_{\Delta\to 0}\frac{F(x_i+\Delta)-F(x_i)}{\Delta}\Delta=1-\mathcal{P}(x\in X)=0
\end{aligned}
\tag{21}
$$

将 f 换为 h，得同样结果。故得证。

□

该结论说明当数据量趋于无限，则真实空间的函数依概率收敛于真实解，且任何可以保证学习误差为零的学习策略得到的估计函数也以概率收敛于真实解。

更进一步地由命题2.3可以得到推论：

推论 2.4. *若设 $(X,\mathcal{A},P)$ 是一个概率测度空间，且其累积概率分布 $F(x)$ 是 Lipchitz 的，即*

$$\forall x,x'\in X,|F(x)-F(x')|\leq L|x-x'|,L<\infty \tag{22}$$

且 D_m 中元素各不相同，则如下等式以概率 1 成立：

$$\lim_{m\to\infty}\mathcal{P}(f(x)\neq h(x)|D_m)=0, \tag{23}$$

证明. 由 $f\neq f^*\Leftrightarrow|f-f^*|>0$，与 $h\neq f^*\Leftrightarrow|h-f^*|>0$，且

$$|f-h|\leq|f-f^*|+|h-f^*| \tag{24}$$

则由命题2.3知：

$$
\begin{aligned}
&\mathcal{P}(|f-h|>0|D_m)\leq\mathcal{P}(|f-f^*|>0|D_m)+\mathcal{P}(|h-f^*|>0|D_m)\\
\Longrightarrow&\lim_{m\to\infty}\mathcal{P}(|f-h|>0|D_m)\leq\lim_{m\to\infty}\mathcal{P}(|f-f^*|>0|D_m)+\mathcal{P}(|h-f^*|>0|D_m)\leq 0
\end{aligned}
\tag{25}
$$

故得证。 □

该结论说明当数据量趋于无限，则任何可以保证学习误差为零的学习策略得到的估计函数都依概率于真实空间的函数相等。

在此基础上我们可以得到：

定理 2.5. *设 C 为学习损失（或效用）；数据的集合为 $D_m=\{(x_1,y_1),(x_2,y_2),...(x_m,y_m)\}$，其中 x_i 为输入，y_i 为***确定性***输出；真实函数空间的表示输入输出关系的函数为 f，其服从以 $P(f)$ 为概率的离散分布，解空间的表示输入输出关系的函数为 h，其服从以 $P(h)$ 为概率的离散分布，h,f 满足确定性输出假设，且 NFL2.1中 $\mathcal{E}(C|f,h,D_m)<\infty$，则以概率 1 下式成立：*

$$\lim_{n\to\infty}E_{hf}(C|D_m)=\lim_{n\to\infty}E_{ff}(C|D_m) \tag{26}$$

其中 $E_{hf}(C|D_m)=E_(C|D_m)$ 是 NFL2.1的左边，$E_{ff}(C|D_m)$ 是 NFL 中令 $h=f$ 得到的期望。

证明. 由推论2.4知下式以概率 1 成立：

$$
\begin{aligned}
&|E_{hf}(C|D_m) - E_{ff}(C|D_m)| \\
=&\left|\sum_{f,h}\mathcal{E}(C|f,h,D_m)p(h|D_m)p(f|D_m) - \sum_{f,f}\mathcal{E}(C|f,f,D_m)p(f|D_m)p(f|D_m)\right| \\
\leq&\left|\sum_{h}\mathcal{E}_h(h)p(h|D_m) - \sum_{f}\mathcal{E}_h(f)p(f|D_m)\right| \\
=&\left|\mathcal{E}_h(h=f^*)p(h=f^*|D_m) + \sum_{h\neq f^*}\mathcal{E}_h(h)p(h|D_m)\right. \\
&\left.-\mathcal{E}_h(h=f^*)p(f=f^*|D_m) - \sum_{f\neq f^*}\mathcal{E}_h(h)p(h|D_m)\right| \\
\leq&|\mathcal{E}_h(h=f^*)p(h=f^*|D_m) - \mathcal{E}_h(h=f^*)p(f=f^*|D_m)| + A + B \\
\leq&\mathcal{E}_h(h=f^*)|p(h=f^*|D_m) - p(f=f^*|D_m)| + A + B
\end{aligned}
\tag{27}
$$

其中

$$
\begin{aligned}
A &= \sum_{h\neq f^*}\mathcal{E}_h(h)p(h|D_m) \leq \left(\sup_h \mathcal{E}_h(h)\right)P(h\neq f^*|D_m) = 0 \\
B &= \sum_{f\neq f^*}\mathcal{E}_h(h)p(h|D_m) \leq \left(\sup_h \mathcal{E}_h(h)\right)P(h\neq f^*|D_m) = 0
\end{aligned}
\tag{28}
$$

故,由于对离散的 $P(f), P(h)$,$\lim_{m\to\infty} P(h=f^*|D_m) = \lim_{m\to\infty} P(f=f^*|D_m) = \lim_{m\to\infty} p(h=f^*|D_m) = \lim_{m\to\infty} p(f=f^*|D_m)$, 故得证。 □

非确定性输出 当输入变量 X 与待预测的输出变量 Y 之间不存在映射关系时，统计学习一般采用最小化经验损失 (ERM) 的方法求得预测函数。这里我们仅考虑如下情形：

定义 2.1. 若对 $((X,Y),\mathcal{A},P)$ 所形成的概率测度空间，以及定义在 Y 与预测函数 $f: X\mapsto Y, f\in\mathcal{F}$ 上的损失函数 $l(y,f,Df,...)$（其中 $Df...$ 为 f 的各阶导数），存在唯一的 f^* 使得 $\forall f\neq f^*$：

$$
\mathcal{L}_{X,Y}(f^*) = \int_{X,Y} l(y,f^*,Df^*,...)dP(X,Y) < \int_{X,Y} l(y,f,Df,...)dP(X,Y) = \mathcal{L}_{X,Y}(f) \tag{29}
$$

则称 $l(y,f,Df,...)$ 及 $\mathcal{L}_{X,Y}(f)$ 为常规损失函数 (regular loss function), 及常规损失泛函 (regular loss functional)。

显然，若 $\mathcal{L}_{X,Y}(f)$ 是严格凸的，则其为常规损失泛函，$l(y,f,Df,...)$ 为常规损失函数。但是一个常规损失泛函不一定是严格凸的。

例 2.1. 比如对 $l(y,f) = (y-f(x))^2$，我们有

$$
\mathbb{E}_{X,Y}(l(f)) = \int_{X,Y}(y-f(x))^2 dP(X,Y) \tag{30}
$$

易知对任意 $0 < \alpha < 1$，

$$
\begin{aligned}
\mathbb{E}_{X,Y}(l(\alpha f_1 + (1-\alpha) f_2)) &= \int_{X,Y} (y - \alpha f_1(x) + (1-\alpha) f_2(x))^2 dP(X,Y) \\
&\leq \int_{X,Y} \alpha^2 (y - f_1(x))^2 + (1-\alpha)^2 (y - f_2(x))^2 dP(X,Y) \\
&< \int_{X,Y} \alpha (y - f_1(x))^2 + (1-\alpha)(y - f_2(x))^2 dP(X,Y) \\
&= \alpha \mathbb{E}_{X,Y}(l(f_1)) + (1-\alpha)\mathbb{E}_{X,Y}(l(f_2))
\end{aligned} \tag{31}
$$

故其为严格凸损失泛函，有且仅有一个最优解。对 $\mathbb{E}_{X,Y}(l(f))$ 求导得：

$$
f^* = \int_X y dP(Y|X) = \arg\min_f \int_{X,Y} (y - f(x))^2 dP(X,Y) \tag{32}
$$

对于常规损失泛函，结合 ERM 的一致收敛性质?? 我们有如下结论：

命题 2.6. 若对 $((X,Y), \mathcal{A}, P)$ 所形成的概率测度空间，以及定义在 Y 与预测函数 $f : X \mapsto Y, f \in \mathcal{F}$ 上的损失函数 $l(y, f, Df, ...)$（其中 $Df...$ 为 f 的各阶导数）有 $\forall \epsilon > 0$

$$
\lim_{n\to\infty} \mathcal{P}(\sup_{f\in\mathcal{F}} |\mathcal{L}^n_{X,Y}(f) - \mathcal{L}_{X,Y}(f)| > \epsilon) = 0 \tag{33}
$$

其中 $\mathcal{L}^n_{X,Y}(f) = \frac{1}{n}\sum_{i=1}^n l(y_i, f(x_i), Df(x_i), ...)$ 为损失泛函 $\mathcal{L}_{X,Y}(f)$ 的经验估计。

令 $h^* = \arg\min_{f\in\mathcal{F}} \mathcal{L}^n_{X,Y}(f), f^* = \arg\min_{f\in\mathcal{F}} \mathcal{L}_{X,Y}(f)$，则

$$
\lim_{m\to\infty} \mathcal{P}(h^* \neq f^* | D_m) = 0 \tag{34}
$$

其中 $D_m = \{(x_1, y_1), (x_2, y_x), ..., (x_m, x_m)\}$ 为样本集。

证明. 令 $h^* \neq f^*$，则

$$
\begin{aligned}
\mathcal{L}^n_{X,Y}(f^*) - \mathcal{L}^n_{X,Y}(h^*) = \mathcal{L}^n_{X,Y}(f^*) - \mathcal{L}_{X,Y}(f^*) \\
+ \mathcal{L}_{X,Y}(f^*) - \mathcal{L}_{X,Y}(h^*) \\
+ \mathcal{L}_{X,Y}(h^*) - \mathcal{L}^n_{X,Y}(h^*)
\end{aligned} \tag{35}
$$

令

$$
\mathcal{L}_{X,Y}(f^*) - \mathcal{L}_{X,Y}(h^*) = -3\epsilon \tag{36}
$$

注意到 $n \to \infty$ 时有

$$
|\mathcal{L}^n_{X,Y}(f^*) - \mathcal{L}_{X,Y}(f^*)| \leq \sup_{f\in\mathcal{F}} |\mathcal{L}^n_{X,Y}(f) - \mathcal{L}_{X,Y}(f)| \leq \epsilon \tag{37}
$$

$$
|\mathcal{L}_{X,Y}(h^*) - \mathcal{L}^n_{X,Y}(h^*)| \leq \sup_{f\in\mathcal{F}} |\mathcal{L}^n_{X,Y}(f) - \mathcal{L}_{X,Y}(f)| \leq \epsilon \tag{38}
$$

以概率 1 成立。故 $\exists \epsilon > 0$ 使得

$$
\mathcal{P}(\mathcal{L}^n_{X,Y}(f^*) - \mathcal{L}^n_{X,Y}(h^*) \leq -3\epsilon + 2\epsilon = -\epsilon) = 1 \tag{39}
$$

故

$$
\mathcal{P}(\mathcal{L}^n_{X,Y}(h^*) > \mathcal{L}^n_{X,Y}(f^*)) \geq \mathcal{P}(\mathcal{L}^n_{X,Y}(h^*) > \mathcal{L}^n_{X,Y}(f^*) + \epsilon) = 1 \tag{40}
$$

由于对任意 $h^* \neq f^*$，有

$$
\mathcal{L}^n_{X,Y}(h^*) \leq \mathcal{L}^n_{X,Y}(f^*) \tag{41}
$$

即

$$\mathcal{P}(\mathcal{L}_{X,Y}^n(h^*) > \mathcal{L}_{X,Y}^n(f^*)) \leq \mathcal{P}(\mathcal{L}_{X,Y}^n(h^*) \geq \mathcal{L}_{X,Y}^n(f^*)) = 0 \tag{42}$$

故与上式矛盾。故当 $n \to \infty$ 时 $h^* = f^*$ 恒成立，令 $m = n$ 得证。 □

类似定理2.5, 我们可以得到：

定理 2.7. *设 C 为学习损失（或效用）；数据的集合为 $D_m = \{(x_1, y_1), (x_2, y_2), ...(x_m, y_m)\}$，其中 x_i 为输入，y_i 为非确定性输出；真实函数空间的表示输入输出关系的函数为 f，其服从以 $P(f)$ 为概率的离散分布，解空间的表示输入输出关系的函数为 h，其服从以 $P(h)$ 为概率的离散分布，h, f 由 ERM 求得，且 NFL2.1中 $\mathcal{E}(C|f, h, D_m) < \infty$，则：*

$$\lim_{n\to\infty} E_{hf}(C|D_m) = \lim_{n\to\infty} E_{ff}(C|D_m) \tag{43}$$

其中 $E_{hf}(C|D_m) = E_(C|D_m)$ 是 NFL2.1的左边，$E_{ff}(C|D_m)$ 是 NFL 中令 $h = f$ 得到的期望。

证明. 该证明与定理2.5的证明基本相同，故省略。不同之处是与推论2.4不同，由于命题2.6恒成立，故该结论也恒成立。 □

3 结论

本文分析了 NFL 定理的极限情况，得到了当样本量趋于无穷大时 NFL 与假设空间的具体选择无关的结论。此前由于 NFL 定理，人们认为 ERM 的学习系统是不能得到真正的最优解的。本文在一定程度上修正了这个认识。此项工作中的分析依赖于分布的一致收敛性（即 Glivenko-Cantelli 定理）的一种局部形式。在此基础上本文得到分布的总变差一致收敛性。此前的结论是经验分布的总变差的一致收敛性是不存在的。这个结论的证明有对大样本统计、数据科学以及人工智能等领域等依赖于海量数据的科学及工程领域有一定的建设性。

参考文献

茆诗松, 王静龙, 濮晓龙, 2006. 高等数理统计. 第 2 版 [M]. 中国: 高等教育出版社.

韦来生, 2008. 数理统计 (中国科学技术大学数学教学丛书)[M]. 中国: 科学出版社.

CANTELLI F P, 1933. Sulla determinazione empirica della leggi di probabilita[J]. Giorn. Ist. Ital. Attuari, 4: 421-424.

DEVROYE L, GYÖRFI L, 1990. No empirical probability measure can converge in the total variation sense for all distributions[J/OL]. Annals of Statistics, 18: 1496-1499. https://api.semanticscholar.org/CorpusID:15581123.

DVORETZKY A, KIEFER J, WOLFOWITZ J, 1956. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator[J/OL]. Annals of Mathematical Statistics, 27: 642-669. https://api.semanticscholar.org/CorpusID:122299729.

VI G, 1933. Sulla determinazione empirica della leggi di probabilita[J]. Giorn. Ist. Ital. Attuari, 4: 92-99.

WOLPERT D H, 1992. On the connection between in-sample testing and generalization error[J/OL]. Complex Syst., 6. https://api.semanticscholar.org/CorpusID:13901468.

WOLPERT D H, 1996. The lack of a priori distinctions between learning algorithms[J/OL]. Neural Computation, 8: 1341-1390. https://api.semanticscholar.org/CorpusID:207609360.

WOLPERT D H, 2002. The supervised learning no-free-lunch theorems[M/OL]. London: Springer London: 25-42. https://doi.org/10.1007/978-1-4471-0123-9_3.